FINITE MIXTURE LATENT TRAJECTORY MODELS

WITH APPLICATIONS TO DEMENTIA STUDIES

Dongbing Lai[1,2], Huiping Xu[2], Barry Katz[2], Daniel Koller[1], Tatiana Foroud[1], Sujuan Gao[2*]

1. Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana
2. Department of Biostatistics, Indiana University School of Public Health and School of Medicine, Indianapolis, Indiana

* Correspondence to:

Sujuan Gao, Ph.D.

Department of Biostatistics

Indiana University Richard M. Fairbank School of Public Health and School of Medicine

410 West 10th Street, Suite 3000

Indianapolis, IN 46202

This is the author's manuscript of the article published in final edited form as:

Lai, D., Xu, H., Koller, D., Foroud, T., & Gao, S. (2016). A multivariate finite mixture latent trajectory model with application to dementia studies. Journal of Applied Statistics, 1–21. http://doi.org/10.1080/02664763.2016.1141181

ABSTRACT (up to 250 words):

We proposed a finite mixture latent trajectory model, which we call it multi-variable multi-domain model. This model can be used to analyze data measured at several domains with multiple tests in each domain to differential the unobserved sub-groups. Such data are usually seen in aging studies, in which different subtype of dementia can only be determined by considering several domains simultaneously. Those tests within each domain are highly correlated and can be considered as measurements of the same underlying latent process from different aspects; therefore, no single test is desirable and we have to model them collectively. Our model can fully utilize all information from all tests while makes the number of parameters estimated at a manageable level. We applied our model to Uniform Data Set (UDS) from National Alzheimer's Coordinating Center (NACC) and 4 groups that have distinct cognitive decline pattern were identified.

KEY WORDS (up to 6):

Finite mixture; latent trajectory; multi-domain multi-variable; dementia

1. Introduction:

Dementia is common in the elderly population and is characterized by the impairment of cognitive function. Patients diagnosed with dementia also present with heterogeneous subtypes including Alzheimer's disease (AD), vascular dementia (VD), Frontotemporal dementia (FTD), and Lewy Body Dementia (LBD) according to different disease etiology. However, accurate differentiation between dementia subtypes is often defined by pathological finding after patients' death, making it difficult to deliver accurate clinical diagnosis of dementia subtype for therapeutic intervention and for scientific investigations. Previous studies have demonstrated differences in dementia profiles by various dementia subtypes with AD patients having dominant problems in memory ((Martin *et al*., 1986; Neary *et al*., 1986; Galton *et al*., 2000) and FTD patients showing more deficit in executive function ((Neary *et al*., 1988, 1998, 2005; Miller *et al*., 1991; Perry and Hodges, 2000). It is of considerable scientific interest to develop a scientific method for accurate clinical subtyping use patients' longitudinal cognitive profiles.

The motivating data set for this study came from the Uniform Data Set (UDS) from the National Alzheimer's Coordinating Center (NACC) with data collected from 30 Alzheimer's Disease Centers (ADC) around the country. The implementation of the UDS started in 2005 where individuals enrolled in ADCs, both dementia patients and cognitively normal controls, agree to undergo detailed annual cognitive assessment using identical cognitive batteries [6]. Neuropsychological tests in multiple cognitive domains including memory, executive function and language were collected annually as well as demographic information including gender, education, and ethnicity [6, 7]. For dementia patients, cognitive evaluation in at least two domains have to be conducted [8, 9]. In addition, multiple tests within each cognitive domain were used for the evaluation. In this paper, we focused on examining the longitudinal trajectories of repeated cognitive test scores to determine patient subtypes using finite mixture latent trajectory models.

Two methods currently dominate the literature on finite mixture latent trajectory models: the group-based trajectory modeling (GBTM) developed by Nagin and colleagues [10-12] and growth mixture modeling (GMM) developed by Muthén et al [13-15]. GBTM, also known as latent class growth analysis (LCGA) [14], assumes that subjects belong to one of several subpopulations/groups with each characterized by unique longitudinal trajectories represented by polynomial functions of time. A key assumption of GBTM is that of conditional independence, *i.e.* longitudinal measures across time within a trajectory group are assumed to be independent. Group membership in GBTM is determined by a multinomial logit model where covariates can be included to investigate their relationship with group membership. On the other hand, GMM accommodates correlations among longitudinal variables across time within a trajectory group by including random effects for each group's trajectory. Therefore, GBTM can be considered as a special case of GMM with no random effect [14]. Both GBTM and GMM assume one trajectory for each longitudinal variable. Therefore, these methods cannot be directly applied to the dementia data described above because it is likely that multiple tests within a cognitive

domain would follow the same trajectory. Proust-Lima et al extended the GMM to jointly model several longitudinal variables that collectively defines a common latent trajectory [2, 16]. However, their models considered only one trajectory pattern per group, *i.e.* within each latent group all tests follow one trajectory no matter which domain they belong to. Therefore, the models are limited to data with a single domain.

In this paper, we propose a finite mixture latent trajectory model to identify cognitive change patterns that can potentially be used for dementia subtyping. In our model, we consider the general setting of multiple latent trajectories with multiple cognitive domains, each of which may contain multiple cognitive tests, a multi-variable multi-domain model. In Section 2, we introduce notations and model set-up. In Section 3 we present parameter estimation procedures using the EM algorithm. We present results from simulation studies in Section 4. Section 5 includes results from the application of our model to the Uniform Data Set (UDS) from National Alzheimer's Coordinating Center (NACC). We present a discussion and conclude the paper in section 5.

2. A Multi-variable Multi-domain Latent Trajectory Model:

In this model, we assume that there are G latent groups. For individual $i,\ i = 1, \dots, N$, we define a G-dimentional vector ω_i, where $\omega_{ig} = 1$ if individual i belongs to group g and 0 otherwise. Suppose there are D domains and K tests. Let y_i be the vector of all observations for individual i, where $y_i^T = (y_{i1}^T, \dots, y_{ik}^T, \dots, y_{iK}^T)$ and n_{ik} be the number of observations for individual i at test k, $k = 1, \dots K$, then each y_{ik} has length n_{ik}. The measurement model if individual i is in latent group g is:

$$y_{i|\omega_{ig}=1} = \Lambda_{i|\omega_{ig}=1}(t_i) + V_i c_i + \varepsilon_i, \tag{1}$$

Where the latent trajectory is defined as:

$$\Lambda_{i|\omega_{ig}=1}(t_i) = X(t)_{1i}\beta_g + Z(t)_i b_{ig}, \tag{2}$$

c_i in (1) is the vector of random intercept for all tests with length K.It is not domain specific and inducing correlation among all test scores from the same individual. Here we assume c_i is distributed as $N(\mathbf{0}, \Sigma_c)$, where Σ_c is a diagonal matrix with σ_{ck}^2 in its diagonal. Design matrix V_i in (1) is $\sum_{k=1}^{K} n_{ik} \times K$ block matrix with the following structure:

$$V_i = \begin{bmatrix} \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}$$

where $\mathbf{1}$ is a column vector of 1s. In k^{th} column, the column vector of 1s has length n_{ik}. ε_i in (1) is an vector of random error with distribution $N(\mathbf{0}, \Sigma_\varepsilon)$, where Σ_ε is a block matrix with $\sigma^2_{\varepsilon k}\boldsymbol{I}_{n_k}$ at diagonal and all other entries are 0s.

$\beta_g$is group specific fixed effects from all domains in group g. Its length is $P \times D$, where P is the number of covariates. Accordingly, covariate matrix $X(t)_{1i}$ has following structure:

$$X(t)_{1i} = \begin{bmatrix} X_{1i1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & X_{1iD} \end{bmatrix}$$

Each X_{1id} has all covariates for all tests in domain $d, d = 1, \dots, D$ with dimension $n_{ik} \times P$. b_{ig} in (2) is group specific random effects for all domains in group g. Similar to β_g, it has length $q \times D$, where q is the number of random effects. It has a multivariate normal distribution $N(\mathbf{0}, W_g^2\boldsymbol{B})$ with $W_1^2 = 1$ and $\boldsymbol{B}$ is the covariance matrix of first group, similarly defined as Proust-Lima et al [2]. Accordingly, the design matrxi $Z(t)_i$ has the following structure:

$$Z(t)_i = \begin{bmatrix} Z_{i1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & Z_{iD} \end{bmatrix}$$

where Z_{id} usually is a matrix of time polynomial of degree $q - 1$ with dimension $n_{ik} \times q$. For example, if $n_{ik} = 3$, for a quadratic model, each Z_{id}has structure like following:

$$Z_{id} = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 \\ 1 & t_{i2} & t_{i2}^2 \\ 1 & t_{i3} & t_{i2}^2 \end{bmatrix}$$

We assume b_{ig}, c_i and ε_i are mutually independent.

For individual i, $i = 1, \dots, N$, the probability that this individual belongs to the latent group $g, g = 1, \dots, G$, is π_{ig}, with $\sum_{g=1}^{G} \pi_{ig} = 1$,This can be modeled through a multinomial logistic regression as:

$$\pi_{ig} = P(\omega_{ig} = 1 | X_{2i}^T) = \frac{\exp(X_{2i}^T \gamma_g)}{1 + \sum_{h=2}^{G} \exp(X_{2i}^T \gamma_h)}, \quad (3)$$

where $\gamma_g$is the group-specific regression coefficients. For identifiability purpose, γ_1 are all set to 0s. Covariates X_{2i}^T used here can be same or different from X_{1i}.

3. Parameter Estimation:

3.1. The EM algorithm:

Since the latent group memberships are unobserved and there are also multiple random effects, the EM algorithm can be used for obtaining parameter estimates [17-19]. Let $\Psi = (\beta_1, \dots, \beta_g, \dots, \beta_G, W_2^{2T}, \dots, W_g^{2T}, \boldsymbol{B}, \Sigma_c, \Sigma_\varepsilon, \gamma_2, \dots, \gamma_g, \dots, \gamma_G)$ be the parameters to be estimated, $f_{ig}(y_i)$ be the density function of y_i in group g, then the observed likelihood is:

$$L(\Psi) = \prod_{i=1}^{N} \sum_{g=1}^{G} \pi_{ig} f_{ig}(y_i) \tag{4}$$

$f_{ig}(y_i)$ has distribution $N(X(t)_{1i}\beta_g, \Sigma_{ig})$, where $\Sigma_{ig} = Z(t)_i W_g^2 \boldsymbol{B} Z(t)_i^T + V_i \Sigma_c V_i^T + \Sigma_\varepsilon$.

Now the unobserved data are $(\omega_i, b_{i1} \dots, b_{ig}, \dots, b_{iG}, c_i)$ and complete data is $y_i^c = (y_i, \omega_i, b_{i1} \dots, b_{ig}, \dots, b_{iG}, c_i)^T$. Then the complete likelihood is:

$$L^c(\Psi) = \prod_{i=1}^{N} \prod_{g=1}^{G} \{\pi_{ig} f_{(y|bc)ig}(y_i | b_{ig}, c_i) f_b(b_{ig}) f_c(c_i)\}^{\omega_{ig}}$$

and:

$$\log(L^c(\Psi)) = \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} \{\log(\pi_{ig}) + \log\left(f_{(y|bc)ig}(y_i | b_{ig}, c_i)\right) + \log\left(f_b(b_{ig})\right) + \log(f_c(c_i))\}$$

$$= \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} \log(\pi_{ig}) - \frac{\sum_{k=1}^{K} n_{ik} + l + K}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} \log|\Sigma_\varepsilon|$$

$$- \frac{1}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} (y_i - X(t)_{1i}\beta_g - Z(t)_i b_{ig} - V_i c_i)^T \Sigma_\varepsilon^{-1} (y_i - X(t)_{1i}\beta_g - Z(t)_i b_{ig} - V_i c_i)$$

$$- \frac{1}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} \log|\boldsymbol{B}| - \frac{1}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} l * \omega_{ig} \log(W_g^2) - \frac{1}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} b_{ig}^T (W_g^2 \boldsymbol{B})^{-1} b_{ig}$$

$$- \frac{1}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} \log|\Sigma_c| - \frac{1}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} c_i^T \Sigma_c^{-1} c_i \tag{5}$$

where l is the dimension of square matrix $\boldsymbol{B}$. From (5), in E step at k^{th} iteration, we need to calculate: $E_{\Psi(k)}(\omega_{ig}|y_i)$; $E_{\Psi(k)}(\omega_{ig} b_{ig}|y_i)$; $E_{\Psi(k)}(\omega_{ig} b_{ig}^T b_{ig}|y_i)$;

$$E_{\Psi(k)}(\omega_{ig} c_i|y_i);\ E_{\Psi(k)}(\omega_{ig} c_i^T c_i|y_i);\ E_{\Psi(k)}(\omega_{ig} b_{ig}^T c_i|y_i).$$

The first one is straight forward: $E_{\Psi(k)}(\omega_{ig}|y_i) = \Pr(\omega_{ig} = 1|y_i)$

$$= \frac{\pi_{ig} f_{ig}(y_i)}{\sum_{h=1}^{G} \pi_{ih} f_{ih}(y_i)}$$
$$= \tau_{ig}^{(k)}$$

(6)

which is the posterior probability of sample i belonging to group g calculated by using the current parameter estimate.

In addition,

$$\begin{aligned} E_{\Psi^{(k)}}(\omega_{ig} b_{ig} | y_i) &= E_{\Psi^{(k)}}(b_{ig} | y_i, \omega_{ig} = 1) \\ &= E_{\Psi^{(k)}}(b_{ig} | y_i, \omega_{ig} = 1) \Pr(\omega_{ig} = 1 | y_i) \\ &= \tau_{ig}^{(k)} E_{\Psi^{(k)}}(b_{ig} | y_i, \omega_{ig} = 1) \end{aligned}$$

Therefore, similarly, we only need to calculate:
$E_{\Psi^{(k)}}(b_{ig}|y_i);\ E_{\Psi^{(k)}}(b_{ig}^T b_{ig}|y_i);\ E_{\Psi^{(k)}}(c_i|y_i);\ E_{\Psi^{(k)}}(c_i^T c_i|y_i);\ E_{\Psi^{(k)}}(b_{ig}^T c_i|y_i)$.

For other conditional expectations, the joint distribution of $(y_i, b_{ig}, c_i)^T$ will be used. It is multi-variable normal distribution with mean:

$$\begin{bmatrix} X(t)_{1i}\beta_g \\ \mathbf{0} \end{bmatrix},$$

and variance:

$$\begin{bmatrix} Z(t)_i W_g^2 \boldsymbol{B} Z(t)_i^T + V_i \Sigma_C V_i^T + \Sigma_e & \Sigma_{ibcg} \\ \Sigma_{ibcg}^T & \Sigma_{bcg} \end{bmatrix},$$

where $\Sigma_{ibcg} = \begin{bmatrix} Z_i W_g^T \boldsymbol{B} & V_i \Sigma_c \end{bmatrix}$ and $\Sigma_{bcg} = \begin{bmatrix} W_g^2 \boldsymbol{B} & \mathbf{0} \\ \mathbf{0} & \Sigma_c \end{bmatrix}$. b_{ig} and c_i are joint together by writing this way. All other conditional expectations then can be obtained by using properties of conditional distribution of multivariate normal distributions. Since:

$$E_{\Psi^{(k)}}(b_{ig,} c_i | y_i) = \Sigma_{ibcg}^T \left(Z(t)_i W_g^2 \boldsymbol{B} Z(t)_i^T + V_i \Sigma_C V_i^T + \Sigma_e \right)^{-1} (y_i - X(t)_{1i} \beta_g) \quad (7)$$

and:

$$\begin{aligned} var_{\Psi^{(k)}}(b_{ig,} c_i | y_i) &= \Sigma_{bcg} - \Sigma_{ibcg}^T \left(Z(t)_i W_g^2 \boldsymbol{B} Z(t)_i^T + V_i \Sigma_c V_i^T + \Sigma_\varepsilon \right)^{-1} \Sigma_{ibcg} \\ &= E_{\Psi^{(k)}}(bc_{ig} bc_{ig}^T | y_i) - bc_{ig}^{(k)} bc_{ig}^{(k)^T} \end{aligned}$$

(8)

From above:

$$E_{\Psi^{(k)}}\left(bc_{ig}bc_{ig}^{T}|y_i\right) = E\begin{bmatrix} bb_{ig}^{T}|y_i & bc_{ig}^{T}|y_i \\ cb_{ig}^{T}|y_i & cc_{ig}^{T}|y_i \end{bmatrix}$$
$$= \Sigma_{bcg} - \Sigma_{ibcg}^{T}\left(Z(t)_i W_g^2 \boldsymbol{B} Z_i^T + V_i \Sigma_c V_i^T + \Sigma_\varepsilon\right)^{-1}\Sigma_{ibcg}$$
$$+bc_{ig}^{(k)}bc_{ig}^{(k)^T} \tag{9}$$

Thus, all conditional expectations can be obtained from (9).

For $\gamma_g^{(k+1)}$ in the model for $\tau_{ig}^{(k)}$, there is no closed form solution, thus it has to be solved numerically. For all other parameters, closed form solutions are available and are given below:

$$\sigma_\varepsilon^{2(k+1)} = \frac{\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{j=1}^{n_{ik}}\tau_{ig}^{(k)}\left(y_{ij} - X(t)_{1ij}\beta_g^{(k)} - Z(t)_{ij}b_{ig}^{(k)} - V_{ij}c_{ij}^{(k)}\right)^2}{\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{j=1}^{n_{ik}}\tau_{ig}^{(k)}}$$

Where $\sigma_\varepsilon^{2(k+1)}$ is the diagonal part of Σ_ε.

$$\sigma_{ck}^{2(k+1)} = \frac{\sum_{i=1}^{N}\sum_{g=1}^{G}\tau_{ig}^{(k)}\left(c_{ik}^{(k)}\right)^2}{\sum_{i=1}^{N}\sum_{g=1}^{G}\tau_{ig}^{(k)}}$$

Where $\sigma_{ck}^{2(k+1)}$ is the k^{th} diagonal part of Σ_c.

$$W_g^{2(k+1)} = \frac{\sum_{i=1}^{N}\sum_{g=1}^{G}\tau_{ig}^{(k)}b_{ig}^{(k)^T}\boldsymbol{B}^{(k)^{-1}}b_{ig}^{(k)}}{\sum_{i=1}^{N}\sum_{g=1}^{G}l * \tau_{ig}^{(k)}}$$

Again l is the dimension of matrix $\boldsymbol{B}$.

$$\boldsymbol{B}^{(k+1)} = \frac{\sum_{i=1}^{N}\sum_{i=1}^{G}\tau_{ig}^{(k)}\, b_{ig}^{(k)^T}b_{ig}^{(k)}}{\sum_{i=1}^{N}\sum_{g=1}^{G}\tau_{ig}^{(k)}W_g^{2(k)}}$$

$$\beta_g^{(k+1)} = \left(\sum_{i=1}^{N}\sum_{g=1}^{G}\tau_{ig}^{(k)}X(t)_{1i}^{T}\Sigma_\varepsilon^{(k)^{-1}}X(t)_{1i}\right)^{-1}\sum_{i=1}^{N}\sum_{g=1}^{G}\tau_{ig}^{(k)}X(t)_{1i}^{T}\Sigma_\varepsilon^{(k)^{-1}}(y_i - Z(t)_i b_{ig}^{(k)}$$
$$- V_i c_i^{(k)})$$

The E-step and M-step will be repeated until the difference of observed likelihood converges to 0. SAS will be used in analysis. To facilitate the estimation, a simpler model without random effects is used to find the start points by using SAS PROC NLMIXED. To avoid local maxima, different start points around the estimates of PROC NLMIXED will be used.

3.2 Posterior classification and model selection:

Posterior classification of subjects into latent groups can be achieved using the posterior conditional probability in (6). Subjects are classified in the latent class for which he or she has the highest posterior conditional probability. To evaluate classification errors of the latent group assignment, we calculate a $G \times G$ matrix, which consists of average posterior probabilities for each latent class among samples classified in each group. The diagonal part of this table will be average posterior probabilities of correct classification and all of them being close to 1 indicates a better fit.

One of the biggest challenges in mixture models is how to choose the number of latent groups. There are numerous studies on this topic with no clear consensus [17]. The traditional likelihood test of comparing a k-group model versus a k+1 group model does not directly apply here because the parameters tested are on the boundary of parameter space thus violating regularity conditions required for the likelihood ratio statistic [20, 21]. Lo, Mendell and Rubin proved that the likelihood ratio statistic was a weighted sum of independent chi-squared random variables with one degree of freedom under the null hypothesis of k groups [21]. However, the required conditions were not always met [22]. A study by Nylund, Asparouhov and Muthen compared various methods for identifying the number of latent groups and found that a bootstrap likelihood ratio test (BLRT) proposed by McLachlan and Peel [17] outperformed all other methods with the caveat that the BLRT method requires intensive computational loads [23]. Within the EM framework, classification-based information criterion were developed [17], aiming to improve the performance of popular information criteria (IC) based methods such as Akake's Information Criterion (AIC) [24] and Bayesian Information Criterion (BIC) [25]. Biernacki et al proposed an integrated classification likelihood criterion (ICL-BIC) [17, 26]. However, less than adequate results were obtained by several simulation studies [11, 17] and the ICL-BIC methods are not wildly used.

In practice, most studies used AIC and BIC due to their ease of implementation [11, 12, 14]. However, these methods to have a tendency to overestimate the number of groups [12, 17, 20, 23]. Hence in practice, the number of latent groups determined by AIC or BIC should be used as a suggestion supplemented by background knowledge [11, 12, 14].

4. Simulation Studies

For the simulation study, we focused on parameter estimation and group classification. The first simulation study was to determine whether our method can correctly estimate the parameters, i.e., whether we can find the true trajectories and assign individuals to the correct groups. The second simulation study was to see whether our method can identify the correct number of latent groups. For both purposes, data were simulated with 4 tests from 2 different domains with each domain having 2 tests. For each domain, linear trajectory was assumed. A binary variable and a continuous variable were simulated for domain specific fixed effects. For group specific random effect, both intercept and slope were assumed. Each sample had three observations at different time points, then one or two observations from some randomly selected samples were deleted; therefore, each sample had one to three observations at different time points. One continuous variable was simulated to determine the group memberships.

In the first set of simulations, five scenarios were used with the assumed number of latent groups between 2 and 6. For each scenario, we generated 500 data sets with each data set consists of 1500 subjects. Under each scenario we fitted a latent trajectory model with the true number of latent groups. There are ?? parameters to be estimated from ?? Table 1 shows the average coverage probabilities of parameter estimates and misclassification rates over all 500 data sets for each number of groups. As can be seen, the parameter estimates and identification of the group membership are very accurate. However, when the number of groups increases, the misclassification rate increases dramatically. This is expected since there is more room for classification error as there are more latent classes. In addition, for a fixed samples size, when the number of latent classes increases, the number of samples within each group decreases and standard errors of the estimates of class-specific parameters increase with a larger number of latent classes. Classification error hence increases with less well separated classes.

# of groups	average coverage (range)	misclassification rate
2	94.89% (92.60%-98.80%)	0.001%
3	95.16% (92.80%-99.00%)	2.29%
4	95.61% (91.40%-99.80%)	8.60%
5	95.73% (93.00%-100.00%)	12.29%
6	96.15% (92.60%-100.00%)	13.97%

Table 1: Average coverages and misclassification rates of simulation results.

For the second sets of simulations, we also simulated 500 data sets and each data set had 1500 subjects. We fixed the number of latent groups to 4 and fitted linear models with 2 to 6 groups and calculated BIC. Based on BIC, only 5 data sets chose the wrong number of groups. Figure 1 displays the average BIC of these 500 data sets. The model assuming 4 latent groups had the lowest BIC.

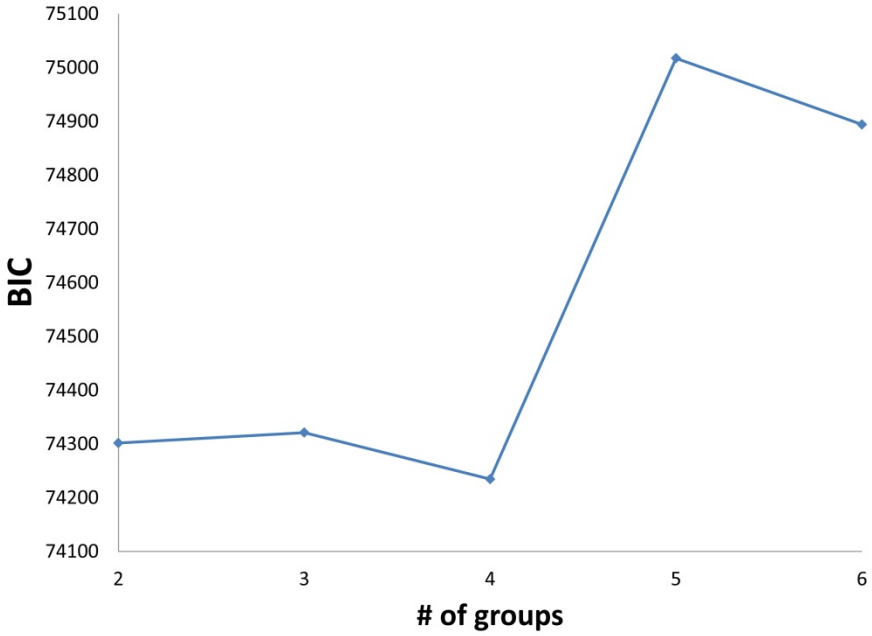


Figure 1: Average BIC of simulated results.

5. Application to the UDS data:

For the purpose of demonstrating our model framework, we used a more homogeneous subset in the UDS by only including Caucasian patients with dementia who had at least four cognitive evaluations. We also restricted analyses to those with cognitive decline after 60 years in order to exclude patients with early onset dementia. Tests from two domains were used: logical memory immediate and delayed recalls tests for the memory domain; Animal Fluency Test and the Boston Naming Test for the language domain. As indicated by Weintraub et al, age of onset, gender and education were included in both the group membership model and the latent trajectory model [7]. Final analysis data set included 30,004 observations from 1517 patients. Since these four test scores have different ranges, all outcomes were rescaled to be between 0 and 10 to have a similar magnitude. In addition, education (in number of years) and age of onset (in years) were rescaled to between 0 and 1. The time variable, age, was measured by decades and centered on the mean age. We tested linear and quadratic trajectories with assumed number of latent groups ranging from 2 to 6. Table 2 is the summary of model results.

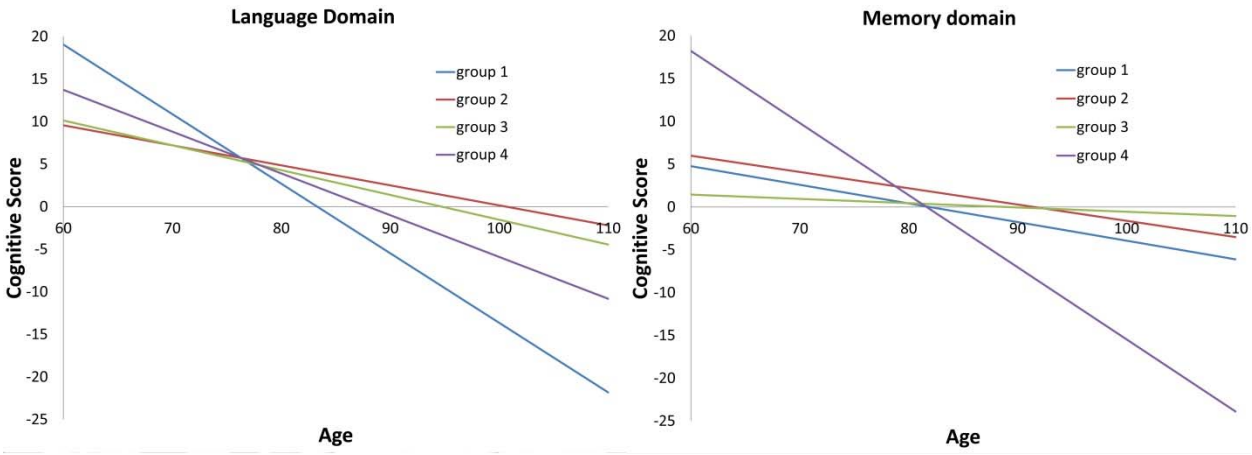
	Linear Model			Quadratic Model		
# of groups	# of parameters	Log Likelihood	BIC	# of parameters	Log Likelihood	BIC
2	43	-47711.57	95738.10	47	-47675.48	95695.22
3	58	-47394.82	95214.45	64	-47329.63	95128.02
4	73	-46992.81	94520.31	81	-46971.34	94535.97

5	88	-46871.30	94387.15	98	-46832.84	94383.48
6	103	-46791.45	94337.31	115	-46713.23	94268.78

Table 2: Log likelihood and BIC of UDS data for different models.

From table 2, the differences between linear and quadratic models are negligible. We chose the linear model for its parsimony and ease of interpretation. Although BIC is always decreasing, it decreases quickly at first and then becomes relatively flat with 4 or more latent classes. Thus, we chose the model with 4 latent groups as the final model.

In figure 2, we plotted model trajectories of a male patient with education and age of onset at the sample means in 4 latent groups using linear model for memory domain and language domain. Group 1 has the steepest decline in language but relatively flat in memory decline; group 4 has the fastest decline in memory and also the second fastest decline in language; patients in groups 2 and 3 have slower decline than those in groups 1 and 4 in both language and memory.



Firgure 2: Estimated trajectories of language (left) and memory (right) decline for male dementia patients in four latent groups.

In Table 3, we present patients' characteristics by the four identified latent groups. For groups 1 and 3, both have more than 70% samples diagnosed as probable AD only. However, compared to group 3, although group 1 has higher percentage of male and more years of education, they have much earlier start age of cognitive decline. This agrees with the observation that AD is a phenotypically heterogeneous disease. The difference between group 1 and 3 can be partially

explained by the role of APOE e4 allele: group 1 has about 73% samples with e4 allele while only 58% in group 3. Several studies already showed that APOE e4 allele is not a risk factor for ~40% of AD [1] and most likely samples used in those studies have similar phenotypes as in group 3, and potentially they share the same etiology. For groups 2 and 4, about half of samples have other type of dementia and not surprisingly, APOE plays a less important role with less than half of samples having APOE e4 allele. However, just as groups 1 and 3, groups 2 and 4 differ dramatically: group 4 has less male, higher education, late age of decline start and more people with APOE e4 allele, therefore, group 2 and 4 may have different etiology.

group	# of patients	male %	Average years of Education (SD)	Average age of onset (SD)	APOE e4 carrier (%)	Probable AD only (%)	Other Dementia (%)
1	300	54.00	15.91(2.89)	70.66(6.56)	72.69	76.00	22.67
2	510	59.80	15.18(3.30)	73.35(7.21)	44.57	41.37	55.88
3	560	46.96	14.16(3.24)	74.03(6.88)	57.79	71.96	27.14
4	147	47.62	16.38(2.10)	76.06(9.00)	47.06	51.02	42.86

Table 3: Characteristics of 4 groups identified.

For this data, we didn't know the true parameters and group memberships; therefore, we used average posterior probabilities to check the model fit. Table 4 is the average posterior probabilities of linear model with 4 groups. Each cell is the average posterior probabilities of all samples classified in that group, for example, for the first row, they are average posterior probabilities in groups 1, 2, 3 and 4 respectively for all samples classified in group 1. The diagonal part is the average probabilities if they are in the "right" group and they are all larger than 0.84, indicating that this model fits our data well.

Classified group	1	2	3	4
1	0.87	0.02	0.05	0.02
2	0.04	0.88	0.10	0.09
3	0.07	0.04	0.84	0.00
4	0.02	0.06	0.00	0.89

Table 4: Average posterior probabilities of 4 groups identified

5. Discussion:

In this paper, we proposed a multi-variable multi-domain latent trajectory model aiming at analyzing data that are often encountered in studies like neuropsychological research. In these studies, there exist latent constructs of multiple domains and within each domain the latent construct is measured by multiple tests. Our model is an extension of GBTM, GMM and model proposed by Proust-Lima et al, and can be used in any studies where more than one test for the same underlying variable is used. We applied our method to UDS data and four distinct cognitive decline patterns were identified.

The extension of handling multiple tests in one single domain is extremely important. In many studies, no single test is desirable and misleading results will be obtained by just selecting one test, therefore, several tests have to be considered simultaneously. The straightforward way to analyze such data is to model each test with a latent trajectory, however, the parameters need to be estimated are enormous and this will cause numerical problem. There exist some methods that aim at reducing dimensionality such as combining tests within the same domain by using sum or weighted average, the downside about these methods are, as indicated by Gray and Brookmeyer, results may not be biological meaningful [27]. In addition, the correlation among the tests is not fully utilized in aforementioned methods [27]. In our method, by modeling tests within the same domain together, the numbers of parameters need to be estimated are greatly reduced. By adding random test specific effects, the difference and correlation among tests are fully accounted. Furthermore, since these tests are measurements of the same underlying latent process from different aspect, model them simultaneously can improve our ability to find the true latent process. The latent process is directly linked to each domain and all results are easy to interpret.

In our model, currently we only assumed normal distribution due to its tractability and easy to implement. This is justified in our analysis of UDS data because although the 4 tests we used are counts of how many items the test subject correctly answered or generated, their ranges are at least 30 and their distributions are not strongly skewed. However, for those tests that have smaller ranges or just binary response, we have to use appropriate distributions accordingly. In the future, we will extend our work to model non-normal variables and/or mixed types variables. Another limitation of using normal distribution lies in selecting the number of groups if we are using information based criterion like BIC. It has been observed in this and many other studies that BIC is always decreasing as more groups are added [11, 17]. This problem is even worse if sample size is big and sample sizes in each group is imbalanced: in those cases, the groups with larger sample sizes can be easily split into two or more groups [11] and currently the best way to address this is using background information as mentioned earlier.

Our UDS samples are all demented and for most of them, data collection were started at the time of their cognitive decline began and they were not in late age of dementia, therefore, our data are not strongly skewed and ceiling and/or flooring effects are not severe. However, this is not a case in many studies. Furthermore, many tests have different sensitivities to patients that are at different stages of dementia [2]. Proust el al proposed a transformation by using cumulative beta distribution, and with only adding 2 more parameters to be estimated, it is very flexible and fits

data well [2]. Jacqmin-Gadda et al proposed a semi-parametric latent process model to better fit the data [28]. However, both methods increase computation complexity. More research is needed in this area.

In this paper, we identified 4 interesting cognitive patterns. However, we didn't clearly link the groups identified to some known subtypes of dementia like dementia with Lewy Boides, Frontotemporal dementia, etc. The reason for this is that this sample consists of mostly AD patients and samples sizes of other diseases are very small. In addition, many patients have two or more subtypes of dementia and the combination of different dementia makes the number of distinct trajectory groups much bigger with each group having only a few samples. UDS data collecting is still ongoing and we will retest our model when we have more known sybtype of dementia samples in future.

Acknowledgment:

The research is supported by National Institutes of Health (NIH) Grants R01 AG019181, and P30 AG10133.

The NACC database is funded by NIA Grant U01 AG016976 (This is required by NACC and we need to submit to NACC first before we send to any journal).

References:

1. *Alzheimer's Disease Genetics Consortium*. Available from: https://alois.med.upenn.edu/adgc/about/overview.html.
2. Proust, C., et al., *A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data.* Biometrics, 2006. **62**(4): p. 1014–1024.
3. Fabrigoule, C., et al., *Cognitive process in preclinical phase of dementia.* Brain, 1998. **121 (Pt 1)**: p. 135-41.
4. Salthouse, T.A., et al., *Interrelations of age, visual acuity, and cognitive functioning.* J Gerontol B Psychol Sci Soc Sci, 1996. **51**(6): p. P317-30.
5. Salmon, D.P. and M.W. Bondi, *Neuropsychological assessment of dementia.* Annu Rev Psychol, 2009. **60**: p. 257-82.
6. Morris, J.C., et al., *The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers.* Alzheimer Dis Assoc Disord, 2006. **20**(4): p. 210-6.
7. Weintraub, S., et al., *The Alzheimer's Disease Centers' Uniform Data Set (UDS): the neuropsychologic test battery.* Alzheimer Dis Assoc Disord, 2009. **23**(2): p. 91-101.
8. McKhann, G.M., et al., *The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease.* Alzheimers Dement, 2011. **7**(3): p. 263-9.
9. Robinson-Whelen, S. and M. Storandt, *Immediate and delayed prose recall among normal and demented adults.* Arch Neurol, 1992. **49**(1): p. 32-4.

10. Nagin, D.S., *Analyzing developmental trajectories: A semiparametric, group-based approach.* Psychol Methods, 1999. **4**: p. 139-57.
11. Nagin, D.S., *Group-based modeling of development*. 2005, Cambridge, Mass: Harvard University Press.
12. Nagin, D.S. and C.L. Odgers, *Group-based trajectory modeling in clinical research.* Annu Rev Clin Psychol, 2010. **6**: p. 109-38.
13. Muthen, B., *Beyond SEM: General latent variable modeling.* Behaviormetrika, 2002. **29**: p. 81–117.
14. Muthen, B., *Latent variable analysis: growth mixture modeling and related techniques for longitudinal data*. The Sage Handbook of Quantitative Methodology for the Social Sciences, ed. D. Kaplan. 2004, Newbury Park, CA. 345–68.
15. Muthen, B. and K. Shedden, *Finite mixture modeling with mixture outcomes using the EM algorithm.* Biometrics, 1999. **55**: p. 463–469.
16. Proust-Lima, C., L. Letenneur, and H. Jacqmin-Gadda, *A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome.* Statistics in Medicine, 2007. **26**: p. 2229–2245.
17. McLachlan, G. and D. Peel, *Finite Mixture Models*. 2004, New York: Wiley-Intersci.
18. McLachlan, G.J. and T. Krishnan, *The EM Algorithm and Extensions*. 1997, New York: John Wiley & Sons, Inc. .
19. Ng, S.K., et al., *A mixture model with random-effects components for clustering correlated gene-expression profiles.* Bioinformatics, 2006. **22**(14): p. 1745-52.
20. Baudry, J., et al., *Combining Mixture Components for Clustering.* J Comput Graph Stat., 2010 June 1. **9**(2): p. 332-353.
21. Lo, Y.T., N.R. Mendell, and D.B. Rubin, *Testing the number of components in a normal mixture.* Biometrika, 2001. **88**(767-78).
22. Jeffries, N., *A Note on 'Testing the Number of Components in a Normal Mixture'.* Biometrika, 2003. **90**: p. 991-994.
23. Nylund, K.L., T. Asparouhov, and B.O. Muthen, *Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study.* Struct. Equ. Model., 2007. **14**: p. 535-69.
24. Akaike, H., *New look at statistical-model identification.* IEEE Trans. Automatic Control, 1974. **19**: p. 716-23.
25. Schwartz, G., *Estimating the dimension of a model.* Ann. Stat., 1978(6): p. 461-64.
26. Biernacki, C., G. Celeux, and G. G, *Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood*. Technical Report No. 3521. 1998, Rhone-Alpes: INRIA.
27. Gray, S.M. and R. Brookmeyer, *Estimating a treatment effect from multidimensional longitudinal data.* Biometrics, 1998. **54**(3): p. 976-88.
28. Jacqmin-Gadda, H., C. Proust-Lima, and H. Amieva, *Semi-parametric latent process model for longitudinal ordinal data: Application to cognitive decline.* Statistics in Medicine, 2010. **29**(26): p. 2723–2731.